Some Open Questions on Multiple-Source Extensions of Adaptive-Survey Design Concepts and Methods

by

Stephanie Coffey, PhD.
U.S. Census Bureau

Jaya Damineni
U.S. Census Bureau

John Eltinge, PhD.
U.S. Census Bureau

Anup Mathur, PhD.
U.S. Census Bureau

Kayla Varela
U.S. Census Bureau

Allison Zotti
U.S. Census Bureau

CES 23-03 **February 2023**

Abstract

Adaptive survey design is a framework for making data-driven decisions about survey data collection operations. This paper discusses open questions related to the extension of adaptive principles and capabilities when capturing data from multiple data sources. Here, the concept of "design" encompasses the focused allocation of resources required for the production of high-quality statistical information in a sustainable and cost-effective way. This conceptual framework leads to a discussion of six groups of issues including: (i) the goals for improvement through adaptation; (ii) the design features that are available for adaptation; (iii) the auxiliary data that may be available for informing adaptation; (iv) the decision rules that could guide adaptation; (v) the necessary systems to operationalize adaptation; and (vi) the quality, cost, and risk profiles of the proposed adaptations (and how to evaluate them). A multiple data source environment creates significant opportunities, but also introduces complexities that are a challenge in the production of high-quality statistical information.

Keyword*:* administrative-records-first and survey-first designs; auxiliary data; field experiments; dimensions of data quality, risk and cost; production systems

[*] The views expressed in this paper are those of the authors and do not reflect the policies of the U.S. Census Bureau. The authors are grateful to Paul Beatty for his helpful thoughts on evaluating administrative data, and to John Abowd, Michael Thieme and Jason Fields for their time and very insightful comments. The authors also thank David Peters for providing Figure 2, and Tamara Adams for providing Figure 3.

1. Introduction

The purpose of survey data collection (and any statistical information collection process) is to generate substantive insight on topics of interest, for a given target population. That insight is often summarized or communicated through statistical parameters of interest at the population or domain level, including means, quantiles, regression coefficients; differences over time or among cross-sectional groups; and measures of variability like variance. The accuracy of these parameter estimates can be reduced by a variety of issues described in the Total Survey Error context (Biemer 2010; Groves and Lyberg, 2010; Biemer et al, 2017), including but not limited to coverage error, sampling error, measurement error, and nonresponse error. Under "ideal conditions", we could accurately specify the types and magnitudes of each of these error sources and incorporate information about error processes into our design plans for sampling, data collection, weighting, and estimation. The current era of increasing data collection costs and reduced rates of unit contact, participation, and wave and item response, however, is far from ideal. In some countries, cost and quality challenges are compounded by the proliferation of surveys, which increases response burden; and by the desire for information at finer levels of granularity, exposing domains where some dimensions of data quality may be particularly problematic.

Survey methodology has evolved to meet multiple changes in survey environments by incorporating new features that respond to the most pressing constraints, which often include the cost associated with high-quality data collection operations. Neyman (1938) developed two-phase sampling to control variances of survey totals by selecting a large sample of units to

collect a small set of information, stratifying on information collected during that first phase, and selecting a small subset of those cases from which to collect detailed information. Additionally, his optimal allocation minimized variance of specified estimators, with respect to cost constraints. Subsampling (Cochran 1977) has long been used in surveys to meet budget constraints while devoting additional resources to a subset of cases to encourage response late in data collection. Dillman (1978) developed the Total Design Method (TDM) as a framework for designing a mail or telephone survey with the "ideal" mix of components. As technologies advanced, new contact and collection modes became possible, allowing modes to be offered simultaneously or sequentially (de Leeuw, 2005) to address different types of error.

Most recently, continued technological advances, including increased computing power and the advent of formal literature on paradata (Couper 2000; 2017), have enabled adaptive and responsive survey designs (Groves and Heeringa 2006; Schouten, Peytchev and Wagner 2017; Tourangeau et al, 2017; Rosenblum et al. 2019). Now, rather than offering a single data collection pathway to all sample units, available data collection features can be tailored to particular sample units to improve the quality and cost profiles for collected survey data.

While continued improvements in paradata capture and real-time data processing will lead to further refinements of the adaptive and responsive design frameworks, survey methodology is at a new crossroads with respect to information production. Administrative data, third-party commercial data, and found data have captured the attention of data users who seek supplementary information that may compensate for declining survey response rates and other survey-quality issues; and who may prioritize low cost, fast production of information that can be provided at finer levels of granularity. Oftentimes, the quality profiles and fitness-for-use of these new data sources are not known *a priori*, because the mechanisms by which these data are

produced do not account fully for all relevant aspects of coverage and measurement (FCSM 2020). However, these issues do not invalidate all uses of such data. Alternatively, information in these new data sources may serve as covariates for predictive models that inform and enhance decisions on sampling, imputation, or adaptive and responsive design. In addition, new data sources may have excellent quality profiles (e.g., coverage) for some domains, leading to potential uses for survey item imputation or supplementation.

The remainder of this paper explores some extensions of adaptive-survey concepts and methods to the capture and integration of both survey and non-survey data sources, which we will refer to as the "survey plus" environment. Some of these extensions focus on adaptation of the data-capture process, and others apply primarily to post-collection processing. Six questions receive principal attention.

(a) What is the general framing for our production of statistical information, and what are the goals for improvement through survey design adaptation?

(b) What are the principal design features we will consider adapting?

(c) What auxiliary data will potentially inform decisions about adaptive changes in the design?

(d) What are prospective decision rules that operationalize auxiliary data and guide specified adaptive design changes?

(e) To what extent will adaptive procedures require, or be enhanced by, systems that capture, integrate and use auxiliary data?

(f) What are the quality, risk, and cost profiles of the proposed adaptive procedures, and what are realistic ways to evaluate those profiles empirically?

Sections 2 through 7 focus on questions (a) through (f), respectively. In each of these sections, we provide examples of how each of these six questions are addressed in the current adaptive survey design environment at the U.S. Census Bureau or in the survey methodological literature, and then discuss how these questions apply to this new environment. Table 1 provides a brief overview of the principal ideas.

Table 1 Prospective Broad Classes of Questions Adaptive Design

Class of Question	**Conceptual Ideas**	**Concrete Examples**
(a) Initial Framing and Goals for Improved Performance	Population(s), estimands, prospective data sources, environmental factors. performance criteria (multiple dimensions of quality, risk and cost), and goals for improvement of those performance criteria.	Goals for improved performance could exist along several dimensions, including: (i) cost (Schouten, Peytchev, and Wagner 2017); (ii) response rate or representativeness (Wagner et al. 2012, Coffey et al. 2020); (iii) Timeliness of data releases (Lohr and Raghunathan; 2017; Elliott and Valliant, 2017; Hand, 2018; Beaumont, 2020), etc.
(b) Levels of Practical Design Decisions	Multi-level structured decisions on each resource type: Funding, data sources, methodology, systems, management, etc.	Usage of web-sourced data needs to be considered across multiple levels, including evaluating: (i) the collection process' ability to meet the legal or regulatory requirements such as Terms of Use; (ii) the consistency and reliability of the web-sourced data; (ten Bosch et al. 2018); (iii) the quality and usability of the data (Mathur, Khaneja, Minoo 2020); and (iv) the coverage for the population of interest.
(c) Prospective Adaptations of Design Features from (b)	For each resource type at a specified level: Adaptive decision options; prospective changes in (conditional) performance profiles (including quality, risk and cost), and conditioning factors that are important for evaluation and improvement of those profiles. Appendix B presents some details of a mathematical development of these ideas.	Adaptive decision rules can allow for features such as stopping work on selected cases (Peytchev 2014, Tolliver 2017), the introduction or withholding of different contact strategies (Coffey et al. 2020), or case prioritization to target where data collection resources should be allocated (Wagner et al. 2012; Tolliver et al. 2019; Dahlhamer 2017; Walejko and Wagner 2018).
(d) Prospective Empirical Information to Inform Adaptive Decisions	(i) "Dispositional" paradata that provide improved information on population-level features, e.g., household or neighborhood characteristics with general explanatory power. (ii) "Situational" paradata that provide improved information on unit-level features. (Silvia, et al. (2013))	There are several types of valuable paradata, including (i) neighborhood-based characteristics, which can be linked to cases by geocoding a random-digit dial sample to attach addresses (Biemer and Peytchev 2012); and case-level contact history paradata, such as the number of household visits by an interviewer, contact strategies employed (e.g., leaving a pamphlet or talking to a neighbor), and type of respondent concerns, such as privacy or burden (Bates, et al. 2010).

(e) Systems of Exploration and Production Steps	(i) Capture and evaluation of production data and related paradata from multiple sources (e.g., surveys, administrative records, web scraping and other forms of organic data) (ii) Implementation of adaptive decisions (iii) Empirical evaluations of components of quality, risk and cost	To implement the selected adaptive design features effectively, a model execution engine must be implemented and connected to the survey data collection ecosystem (Thalji, et al. 2013; Thieme and Mathur 2014), such as the Concurrent Analysis and Estimation System (CAES) that has been implemented by the US Census Bureau.
(f) Evaluation of Quality, Risk, and Cost Profiles of Proposed Adaptive Procedures	Evaluation of approximate performance profiles (including quality, risk and cost), in some cases: (i) Averaging over environmental conditions, or conditional on specific environmental conditions (ii) Treating the full design vector as fixed; or treating some design features as dependent on particular paradata-driven design adaptations	Some adaptive designs may be able to control specific features, such as tailored incentives (Coffey, Reist, and Zotti 2015), leading to evaluations of the marginal impact of the incentive over experimental treatment groups or domains within treatment groups. Other designs, such as case prioritization (Wagner, et al 2012; Tolliver, et al. 2019) or measuring interviewer productivity (Vandenplas, Loosveldt and Beullens 2017) evaluate a general *protocol*, where the impact of case prioritization is the cumulative impact of many individual contact attempts that follow the prioritization protocol..

2. Key Goals for Improvement through Adaptation

In principle, production of statistical information should use a design that balances numerous dimensions of quality and cost. Important quality dimensions generally will include accuracy (as reflected in the mean squared errors of specified estimators, and the widths and coverage rates of interval estimators), as well as other components like comparability, granularity, punctuality, interpretability, accessibility, and relevance. See, e.g., Brackstone (1999) and National Academies (2017). Cost dimensions potentially include all resources allocated to all steps in the statistical production process, e.g., cash; data sources; methodology and technology used for the capture, management, and integration of those data sources; management capabilities; and related personnel (especially those with scarce skill sets). Each cost dimension may include both fixed and variable cost components (Olson, et al. 2021). Some of the variable cost components may be attributed to a specific concrete part of the statistical production process (e.g., interviewer training or travel). Analysis of other variable components may encounter issues with confounding of multiple prospective causes; or with complex dependencies (e.g., due to supervisor judgment-based interventions).

Comprehensive adaptive management of these dimensions of quality and cost generally will be unfeasible due to many issues related to both measurement and control. Consequently, adaptive and responsive design work has tended to focus on specified measurable indicators of quality and cost, e.g., quality measures based on response rates or R-indicators; or cost measures based on the number of contact attempts or interviewer hours. As we consider extensions from customary adaptive surveys to "survey plus" environments, it will be useful to consider a range of prospective quality and cost measures that are computationally feasible and that provide realistic context for adaptive design decisions.

2.1. Current Environment

Adaptive and responsive designs tailor data collection features to impact measures of data quality, such as nonresponse error or measurement error, or cost (Schouten, Peytchev and Wagner 2017). The tailoring decisions are made in pursuit of a pre-defined survey goal, such as increasing response rate or balance in the respondent population (Wagner et al. 2012, Coffey et al. 2020); reducing the variance of key survey estimates or the variation in weighting adjustments (Beaumont et al. 2014; Paiva and Reiter 2017); or controlling data collection costs (Peytchev 2014; Wagner et al. 2021; Coffey and Elliott 2022). As a result, adaptive and responsive designs typically have increased effort (i.e., resources) applied to certain cases. This generally leads to decreased effort applied to other cases, in order to meet budgetary constraints. The tradeoffs then can be explained as "maintaining or improving data quality for a fixed cost" or "reducing cost while maintaining data quality."

Tailoring and interventions are informed by estimated data collection parameters, such as propensity to respond, propensity to have a particular data collection characteristic, predicted values of responses, predicted costs of obtaining a response (or nonresponse), etc. The parameters selected are a direct result of the goals of the adaptive design itself (Groves and Heeringa 2006). If reducing nonresponse error is the goal, R-indicators or coefficients of variation (CVs) of response propensities could be used to guide interventions, while R-indicators or CVs of nonresponse weighting adjustments could be used as evaluation criteria (Schouten, et al. 2009; Schouten, et al. 2011). As a more illustrative example, we can consider the Survey of Income and Program Participation (SIPP). The SIPP is a longitudinal, in-person, and nationally representative survey. In 2017, the SIPP implemented a series of interventions in an attempt to improve both response rates and representativeness. In order to achieve both of these goals, the

survey team developed prioritization rules that would help to redistribute resources to cases that were both likely to respond and would produce a more representative sample (Tolliver, et al 2017). We will expand more on these prioritization rules in the next section.

Alternatively, if the goal of the design is to maximize response rate, response propensity might be used to identify cases for intervention, while response rate and cost could be used as the evaluation criteria. Finally, if the goal is minimizing cost without harming the quality of a key survey statistic, then estimated costs and predictions of survey responses could be used to enact interventions while observed costs and the variance of actual survey responses could be used as evaluation criteria.

2.2. Survey Plus Environment

In this new environment, it will still be important to produce high quality information and control the costs of the information production process. As we move from the current environment to a survey plus environment, we may be able to broaden the set of available options for data quality improvement. Of special note are cases in which one may reduce the aggregate cost and burden of survey data collection through extensive and carefully targeted use of imputation. For example, one may develop an imputation procedure through rigorously developed regression or hierarchical models that use some directly collected survey data in conjunction with extensive amounts of administrative record or other non-survey data. The resulting procedure may lead to satisfactory quality and efficiency, depending on the model goodness-of-fit and the conditional variances of the resulting unit-level imputations. In addition, we note that direct replacement of a survey item with an administrative record value amounts to imputation based on a single-variable regression model with a slope equal to one and an intercept equal to zero. For example, in the 2017 SIPP case prioritization experiment, cases where administrative records were likely

to be available could be de-prioritized. This allowed for the reallocation of resources to other cases where alternative data sources might not be available.

For example, in the current survey environment, improving timeliness may lead to offering multiple response modes to increase participation convenience, increasing the number of interviewers to reduce the length of the data collection period, or restructuring data processing to reduce the time between the end of data collection operations and the release of the survey data for users. In the survey plus environment, increased timeliness of data releases may arise from imputation or the supplementation of collected survey data with administrative data, reducing the number of cases involved in survey operations, or using statistical techniques to integrate multiple data sources (Lohr and Raghunathan; 2017; Elliott and Valliant, 2017; Hand, 2018; Beaumont, 2020) in order to release data more frequently.

Expanded work with survey nonresponse and non-survey data sources have led to further empirical exploration of bias, which in many cases may make the predominant contribution to mean squared error (Meng 2018; Bradley, et al. 2021; Rao 2021). It would be of interest to extend these concepts and methods to adaptive design cases. In particular, if sensitivity analyses and other empirical work indicate that bias may dominate mean squared error of a "survey plus" estimator, then one may wish to explore the ways in which adaptive procedures may help to reduce some important sources of that bias. In addition to goals related to quality and cost, alternative data sources have led to new goals for survey production. The National Academies of Sciences (NAS 2017; 2021) have noted that the interest in alternative data sources stems at least partially from the need to generate estimates at more granular levels, e.g., subdomains defined by relatively fine-level geographical or demographic classifications. While small area estimation techniques are already utilized for this purpose, alternative data sources could increase coverage

at the item or domain level, increasing the information support upon which small area estimation methods could rely. Additionally, the use of alternative data sources may allow us to estimate regression coefficients or relationships between larger sets of covariates or estimates than before, by allowing more varied and voluminous sources of data to be linked than would be possible with a single survey. These richer, more granular data products can help researchers, policymakers and data users better understand characteristics of the population or its business. Additionally, goals may be related to the minimization of respondent burden, which could lead to the use of alternative data sources as a primary choice, only conducting survey data collection for sample units where we do not have alternative data, or where the quality of the alternative data is unacceptable.

It is important to keep in mind that there may also be competing goals that are at odds with more detailed, richer, more granular information, such as protecting the privacy and confidentiality of individuals or businesses in the domains of interest (NAS 2017; 2021). Therefore, there may be additional goals such as minimizing the risk of disclosure of an individual based on information in released data or estimates. Similar to our current environment, adaptive and responsive decision-making often comes down to navigating tradeoffs. In the new survey environment, we may have tradeoffs such as "minimizing the variance of key subdomains while minimizing the likelihood for disclosure." Similar comments also apply to prospective adaptations to manage trade-offs involving other dimensions of quality, e.g., the interpretability and the cross-sectional or temporal comparability of published estimates.

As additional alternative data sources are considered for information production, it will also be increasingly important to consider the principles and practices for federal statistical agencies (Citro 2014b; NAS 2021). For example, the incorporation of administrative data or commercial

third-party data would require appropriate levels of trust by data providers, as well as increased openness about sources and limitations of the information that is released. In general, the new survey environment will lead to a more nuanced discussion of the goals of a particular information production operation, so that the best choice of operational features, auxiliary data, and decision rules are identified to improve the adaptive or responsive approach.

3. Principal Design Features Under Consideration

In work with both customary sample surveys and the integration of multiple data sources, design decisions generally involve multiple levels of decisions on resource allocation, with each level depending on decisions made at previous levels. Consequently, adaptative decisions take place within the context determined by those levels of design decisions. Appendix A provides some illustrative examples based on, respectively, customary stratified multistage sample surveys; appending administrative record data to sample units; use of sample surveys to supplement data capture that is centered primarily on administrative records; and web-scraping.

3.1. Current Environment

Currently adaptive and responsive procedures focus on data collection operations themselves and can be roughly divided into two types: recommendation and deterministic. Recommendation decision rules most frequently occur in decentralized, interviewer-administered operations, like Computer Assisted Personal Interviewing (CAPI), where the interviewer may have unique information about the data collection process, and where there is less centralized control over how workloads are managed. This environment can lead to adaptive procedures that are informational but still give the interviewer discretion on how those procedures are implemented. Here, adaptive procedures are typically a variation on case prioritization where cases are prioritized to decrease the risk of survey error (Wagner et al. 2012), increase the

representativeness of harder-to-reach populations (Tolliver et al. 2017; Tolliver et al. 2019; Dahlhamer 2017), or increase overall response by targeting the cases most likely to respond (Walejko and Wagner 2018).

Deterministic procedures, on the other hand, lead to a specific action, such as introducing a particular data collection feature, such as an incentive or a mailed questionnaire (Coffey et al. 2015; Lavrakas 2018; Jackson 2020); introducing or withholding a mode of data collection in multi-mode surveys (Coffey et al. 2020); or putting cases on hold or removing them from interviewer workloads to reduce effort applied to specific case (Dahlhamer 2017; Tolliver et al. 2017; Wagner et al. 2021; Coffey and Elliott 2022). For example, the 2017 SIPP prioritization made use of both static and dynamic prioritization rules. The set of static rules were put in place to help prioritize individuals that could not be linked to administrative records, as well as individuals who likely moved between interview periods. The set of dynamic, model-based rules aimed to prioritize households that are both the under-represented and likely to respond (Tolliver et al. 2017).

Once we have identified the collection of features that *could* be used for adaptation, relationships between those features and their expected effects on the potential goals outlined in Section 2 will inform what features *should* be considered for adaptation. One obvious, but still important, limitation of adaptive procedures available in the current environment is that the features available for adaptation are constrained by the available set of survey data collection activities as well as how long it takes to actualize an adaptive intervention. For example, some prospective adaptations, such as changing letter wording, mailing questionnaires, or sending incentives, would not be available for surveys only use interviewer-assisted in-person interviewing.

Similarly, the adaptations described above that the SIPP could implement, would not be available for a survey that does not use in-person interviewing.

It is also important to remember that survey operations (and other types of information capture and production) are subject to external factors over which statistical organizations have little to no control, and which can affect the output from the baseline data collection operations, as well as that of any adaptive procedures. For example, Larsen, Lineback, and Reist (2020) found that refusal rates in the Current Population Survey (CPS) were associated with a variety of environmental factors, including the unemployment rate, inflation rate, GDP, the presidential approval rating, whether it is a decennial Census year, and consumer sentiment scores. Survey data collection output also can be affected by advances in technology, such as with the increase in internet access (Callegaro et al. 2015) or the increase in cell phone-only households (Bates 2009; Blumberg 2021). Even major events have impacts on data collection outcomes, such as unanticipated natural disasters or the recent COVID-19 pandemic (BLS 2021). As a result of this uncontrollable variability, adaptive and responsive survey features are often built on top of a relatively static baseline data collection operation.

3.2. Survey Plus Environment

In this new survey environment, we will consider additional sources of data for different design decisions throughout the survey lifecycle. These auxiliary data sources could be used not only to enhance current adaptive procedures, but also to supplement data collection responses. For example, survey procedures (including adaptive features) may be focused primarily on collection of data from subpopulations that are not covered adequately by the available non-survey (e.g., administrative) data. Similarly, some of the survey procedures may center on estimation of the

coefficients of models used to adjust to variable-specification or unit-definition problems; or models used for imputation of missing data, at the item or unit level.

Access to these additional sources of data before the start of data collection could allow survey teams to tailor their sampling strategies. In an experimental setting, improved propensity models could help identify likely respondents across experimental groups (Zotti 2019). These additional sources of data could also be used after data collection to improve imputation and weighting models (Benedetto, et al. 2015; Giefer, et al. 2015).

Depending on the quality and reliability of these auxiliary data sources, we might even consider a range of imputation and modeling options to supplement standard survey operations for specific domains or survey items. If we have an additional data source that reliably provides data on items of interest for a specific domain of cases, we might consider imputations based on these external data. For example, once a sample unit with satisfactory auxiliary data has received a certain number of contact attempts, or has a response propensity below a given threshold, the unit could be removed from data collection procedures to redirect resources (Mule, et al. 2021). This would allow survey teams to implement survey operation stopping rules while losing relatively little information for those ‘stopped’ cases.

Alternatively, some of these additional sources might be reliably available for nearly all cases but might only address one survey item, prompting the potential removal of that item from the survey’s questionnaires/interviews. By removing high-burden questions from survey questionnaires/interviews and instead using reliable auxiliary data sources to produce those item estimates through imputation, we could potentially reduce respondent burden.

In some cases, auxiliary data could function as the primary source of information for an estimate, with survey operations providing supplemental data collection. In this instance, the data collected

through survey operations could be used to correct for gaps in coverage in the auxiliary data source, to construct adjustment models for auxiliary data to account for temporal or measurement issues, or to validate that the auxiliary data is creating estimates that accurately reflect estimates for topics typically measured through survey-first operations. Realistically, however, these new data sources will have limitations of coverage, accuracy, and consistency, and so it is unlikely they will replace survey-based data collection entirely (Cornesse 2020). Further, target variables for many surveys may not be present in alternative data sources. If certain items are not available in alternative data sources, traditional survey data collection operations will continue to be needed, even if only in a limited manner to enable imputation or other model-based methods for estimation. Additionally, even high-quality alternative datasets will need to be linked, either to a population frame or a survey sample, in order to enhance adaptive procedures or provide new ways of producing information. Record linkage (Christen 2019), either through unique identifiers, probabilistic linkage, or even data fusion, is a critical area for research, as it will enable survey data and alternative data sources to be integrated to improve the production of information in a cost- and time-effective manner.

To think seriously about adapting data collection to utilize these various alternative data sources, we first need to consider what types of auxiliary data will be available. These sources might vary dramatically from survey to survey, and possibly even from year to year within a given survey program. Because of these variations, thorough evaluation of potential data sources is required prior to making any significant changes to data collection processes. We will need to outline a system of metrics to evaluate the quality of these new data sources, and the feasibility of incorporating them into data collection procedures. In addition to these evaluation metrics, we will need to consider what systems might be needed to leverage these data sources for

adaptation. Throughout the next two sections, we will further discuss the types of auxiliary data that are up for consideration, as well as the systems we will need in place to effectively and efficiently utilize them.

4. Necessary or Desired Auxiliary Data

4.1.Current Environment

To inform adaptations and tailoring in the current setting, auxiliary data are necessary to predict the parameters that drive adaptation under different scenarios. Survey organizations currently leverage a variety of auxiliary data sources across different operations for adaptive and responsive strategies. Administrative data or commercially available data may be linked to the survey sample to provide more information for stratification and assignment to different contact strategies prior to data collection (Zotti 2019; van Berkel, et al. 2020). For example, the National Teacher and Principal Survey (NTPS) has historically made use of commercially available data. The NTPS is a national, cross-sectional survey of public and private schools. They purchase vendor data in the form of teacher lists, detailing the names and positions of teachers at the schools in sample. In the event that a school does not provide the requested list of teachers, they can pull that information from the vendor data instead. If a school does not have available vendor data, they may be prioritized in data collection (Zotti, 2019).

In addition to data that enrich the frame, paradata (Couper 2000; 2017) are also vital for estimating response propensity, a parameter commonly used to inform interventions during the data collection period (Groves and Heeringa 2006; West, et al. 2021). Predicted or imputed responses, or similar information to that being collected in the survey may be compared to accumulating response data to make intervention decisions about the quality of the incoming response data (Morris, Keller and Clark 2015). Additionally, information about the costs of

survey operations may be used to predict future data collection costs (Wagner 2019; Wagner, et al. 2020) to guide intervention decisions about future data collection features, such as the optimal time to move cases from one phase of data collection to the next (Wagner, et al. 2020).

Given the broad array of auxiliary data, and the importance of these data sources to the estimation of data collection parameters that inform adaptive and responsive interventions, many researchers have investigated the quality of these data sources (Bates, et al. 2010; Biemer and Peytchev 2012; West and Kreuter 2013; Valliant, et al. 2014). As new data collection technologies have emerged and more information becomes available to append to the sampling frame, the sample, or attempt-level information during data collection, research into and analyses of the quality of these data sources and items will continue to be important.

4.2.Survey Plus Environment

In the new environment where survey data will be combined with other existing data, additional information will be needed about features of the administrative, commercial, or found data (alternative data) to allow survey organizations to determine when and how these new data sources could be used for adaptive and responsive procedures. Just as auxiliary data currently are used to predict data collection parameters or understand measures of progress, quality and cost of survey data collections, measures of usability, quality and cost will need to be developed for these new data sources.

For example, in the current environment, methodologists may use estimates of response propensity given different features (e.g., incentives, particular modes), in order to determine which features a particular sample unit should receive. Here, with the availability of alternative data for supplementation, we may need to consider not just the propensity of a case to respond during data collection, but also the quality and cost of the information in the alternative data

source that could be used to take the place of survey response data. As a conceptual example, we may determine that a case with a very high survey response propensity should remain in data collection, while cases with lower survey response propensities should remain in data collection unless the alternative data source passes a high threshold for quality (e.g., high accuracy, very current data). Cases with the lowest response propensities may be eligible for supplementation with information from alternative data sources with a slightly lower threshold for quality (e.g., high accuracy, slightly older data). We would be unlikely to implement this particular adaptive strategy in the current environment (because of the concern about inducing bias by stopping cases with low response propensities), so this represents an expansion of potential intervention options. Yet, consistent with the current adaptive and responsive protocol, the decisions that are made are based on tradeoffs between costs and errors. For example, in 2021 the NTPS began research into the feasibility of using web scraped data in their data collection. A program is being developed to scrape information from school websites about teachers and principal employment. However, it is clear that not all school websites are up-to-date, or are in a format that can be easily scraped (source?). If the NTPS were to consider prioritizing some schools over others, they might consider prioritizing schools that do not have an online presence, or do not have teacher information available online in a format that can be easily collected and processed by their program.

Similar to the way auxiliary data are used in surveys to generate measures of quality, progress or cost, measures of quality and cost could be constructed from auxiliary information about the alternative data sources. Auxiliary items that could inform measures of quality could include: the risk of losing access to the alternative data; the age or recency of the alternative data, especially with respect to the reference period of the desired estimate; the granularity of the data (e.g., is the

data at the sample unit level or at a coarser level); accuracy; and reliability. Measures of cost could be generated from past purchases of data, information from producers, aggregators, or sellers of alternative data, and the reusability of the data (either over time or across projects). The Federal Committee on Statistical Methodology (FCSM) recently published guidance on these types of measures of data quality for alternative data (FCSM 2020). The availability of auxiliary data from which to derive these measures is critical to understanding the cost and quality properties of alternative data sources, helping survey organizations determine how best to leverage them for information production.

5. Prospective Decision Rules

5.1.Current Environment

Current adaptive and responsive frameworks utilize a variety of decision rules to inform procedures and protocols. Decision rules fall into one of several categories: threshold-based, rank-based, or optimization-based. Threshold-based decision rules are straightforward: when a sample unit crosses a threshold, it becomes eligible for an adaptation or intervention. Peytchev (2014) and Tolliver, et al. (2017) used thresholds to stop work on cases with estimated response propensities below a pre-defined value. Coffey, et al. (2020) used a threshold to introduce or withhold contact strategies for sample units based on fixed values of partial R-indicators (Schouten, et al. 2011). Thresholds can be difficult to define strictly, as implementations of a survey do not have exactly the same environmental conditions over time. As a result, thresholds can seem overly rigid. In addition, if there is a large change to data collection progress or costs from one implementation to the next, a rigidly defined threshold can lead to more (or fewer) cases being identified for intervention than the budget or survey requirements can handle.

As a result, many decision rules are rank-based, and are implemented by identifying the "top," "bottom," or "x percentage" of cases for a treatment. This allows survey organizations to employ adaptive and responsive procedures while hedging against too many (or too few) cases receiving an intervention. Coffey et al. (2015), identified the 20% of cases with the highest weighted response influence (Särndal and Lundström, 2008) for incentivization. The percentage of cases was set at 20% to meet budget requirements, which is often a limitation on higher-cost interventions. Dahlhamer (2017) prioritized the top and bottom 25% of open cases as part of an adaptive design experiment to ensure a reasonable number of cases were affected by adaptive procedures. Tolliver et al. (2019) used a rank-based rule to ensure that no fewer than one case, and no more than 20% of cases in any interviewer caseload were assigned high priority, in order to avoid a situation where a case load had too many high priority cases for the interviewer to actually carry out adaptive procedures on all of their high priority cases.

The last category of decision rules rely on optimization, by either maximizing or minimizing some function of survey data collection parameters to identify the optimal set of cases for intervention. Coffey and Elliott (2022) applied cost- and effort-reduction interventions to the set of cases that minimized the product of root mean squared error (RMSE) of the mean of a key survey statistic (salary) and data collection costs. The minimization function was defined as:

$$O^{S_A} = \left(\frac{RMSE\left(\hat{\bar{y}}_t^{S_A}\right)}{RMSE\left(\hat{\bar{y}}_t^{S_0}\right)}\right)\left(\frac{\left(\hat{C}_t^{S_A}\right)}{\left(\hat{C}_t^{S_0}\right)}\right),$$

where alternate strategy $A = \{2, 4, 6, \ldots, 96, 98, 100\}$, based on the percentage of open cases being switched to the alternate strategy at time t; $RMSE\left(\hat{\bar{y}}_t^{S_A}\right)$ is the RMSE of the mean of salary under the alternate strategy A; $RMSE\left(\hat{\bar{y}}_t^{S_0}\right)$ is the RMSE of the mean of salary under the baseline

condition (where no cases are identified for intervention at time t); $\hat{C}_t^{S_A}$ is the estimated total data collection cost under alternate strategy A; and $\hat{C}_t^{S_0}$ is the estimated total data collection cost under the baseline condition (where no cases are identified for intervention). The optimal set of cases to switch to the alternate strategy satisfies $\min_A\{O^{S_0}, O^{S_2}, O^{S_4}, O^{S_6}, \dots, O^{S_A}, \dots, O^{S_{100}}\}$.

Wagner, et al. (2021) also defined an optimization rule for identifying cases for intervention and sought to minimize a function of data collection costs and mean squared error (MSE) of several key survey variables. Optimization-based rules are the most complicated to implement as they often require predictive models for survey estimates, response propensity, and data collection costs under different sets of data collection strategies. However, they provide a statistically rigorous method to identify cases for intervention and their expected effect on cost and quality measures of interest. As shown in these examples, they can also be designed to take into account the effect of interventions on the survey estimates themselves, rather than indicators for quality or bias, such as response rates, CVs of response propensities, R-indicators, or other proxies.

5.2.Survey Plus Environment

In the new environment, all three types of rules will continue to exist, and may be useful in different settings. Again, the rules that are defined should support the information production goal(s). An example of a threshold-based rule would be one where imputations based on alternative data sources are used instead of direct survey data collection for domains where the alternative data sources have at least 80% coverage of the target population and produce imputations with sufficient accuracy. Domains not meeting these thresholds would be subject to standard survey data collection operations.

A rank-based rule could consider using alternative data sources for cases based on a case-level data quality score, $\hat{q}_i$, which could be defined as:

$$\hat{q}_i = \frac{\hat{r}_i}{\hat{\rho}_i \hat{e}_i} \ ,$$

where the quality score, $\hat{q}_i$, for the i^{th} case is a function of $\hat{r}_i$, an estimate of the accuracy of the item from an alternative data source; $\hat{\rho}_i$ is the estimate of response propensity on the next data collection attempt; and $\hat{e}_i$ is an estimate of the accuracy of the item response value if the sample unit responded during data collection.

It is reasonable to imagine a scenario in which the error distribution of the record stays relatively constant, but the propensity of a response on the next contact decreases, and the propensity that the information is ever collected through the survey also decreases. In this scenario, the longer data collection goes on, the larger $\hat{q}_i$ becomes. We could sort open cases based on $\hat{q}_i$ from largest to smallest, with the largest values reflecting significantly more confidence in the information from administrative record data versus the information that might be collected during survey data collection. Imputation rules could be considered for cases with, e.g., the top 20% of $\hat{q}_i$.

An optimization rule might resemble those described above in the current environment, but focus on alternative goals. As an example, we could assume our goal was to minimize effort spent on attempting to interview a set of likely unproductive sample units, such as vacant housing units. We could consider this a minimization problem where we want to identify the set of cases, x, which are part of the set of open housing units, N, that minimizes the function:

$$O^{S_A} = \left(\frac{\hat{B}_t^{S_A}}{\hat{B}_t^{S_0}}\right)\left(\frac{MSPE^{S_A}(\hat{V})}{MSPE^{S_0}(\hat{V})}\right) \ ,$$

where $\hat{B}_t^{S_0}$ is the effort we expect to expend (e.g., in dollars, hours, attempts, etc.) under the standard data collection strategy, and $MSPE^{S_0}(\hat{V})$ is the estimated mean squared prediction error of vacancy status that we would obtain by carrying out the standard data collection strategy, while the $\hat{B}_t^{S_A}$ and $MSPE^{S_A}(\hat{V})$ are the estimates of effort and prediction error we would expect to generate under a strategy where data collection is stopped for some set of units A. For the units where data collection is stopped, units predicted to be vacant are considered to be vacant.

Other functions of effort and error could be constructed. Measures of effort could be estimated using past information from past data collection efforts or expert opinion from data collection staff or other sources (Coffey, et al. 2020). Measures of prediction error could incorporate error associated with linkage error, nonresponse, measurement error, and/or a measure that accounts for confidence in predictions to help identify cases that will have the largest impact on survey goals.

These decisions could be made overall or at finer geographic levels. It is important to keep in mind that, while identifying goals of the information production process and developing decision rules for interventions, the data required to evaluate those business rules must be available and in a usable format at the time of intervention. This can be challenging for more complex decision rules that are applied during a survey data collection operation.

6. Systems for Capture, Integration and Use of Auxiliary Data

6.1.Current Environment

To implement an adaptive design in a multi-survey environment with multi-mode surveys, a model execution engine must be implemented and connected to the survey data collection ecosystem (Thalji, et al. 2013; Thieme and Mathur 2014). The model execution engine draws

intelligence from various inputs received from the data collection ecosystem and is capable of driving interventions that affect various data collection modes. Typical inputs to the model execution engine include: sampling frame data, accumulating survey paradata for all modes, interviewer and case assignment data, response data, cost and effort data, as well as administrative and other third-party data. Interventions based on available data collection features may include: adding a case to a particular mode (e.g., telephone or in-person interviewing), changing the priority of a case, pausing or stopping work on a case, sending or withholding mailings, among others.

Minimalist implementations of model execution engines may have simple file-based interfaces with survey Operational Control Systems (OCS). Input data may be read from files arriving at a set frequency and interventions may be sent to the OCS for action that affects the data collection modes. The models being executed can vary from simple branching logic and mathematical functions to AI-based models.

At the U.S. Census Bureau, the core of the adaptive design ecosystem is the Concurrent Analysis and Estimation System (CAES), a platform that, in near real-time, can run models to inspect responses; perform microdata-level coding and editing; execute statistical models, including imputation and weighting; and produce survey estimates and variances of those estimates, all on a flow basis. CAES statistical models can perform activities such as concurrent analyses to determine the status of individual responses as well as the overall state of data collection for a given survey. The outcome of the analyses can update the OCS case status and trigger actions such as starting a case, stopping a case, or changing case mode assignment, as well as changing the course of planned data collection activities. CAES supports an agreed upon suite of statistical and analytical tools, including software packages such as SAS, R, and Python. The flexibility of

the CAES design allows for the rapid addition of new software packages as needed. Current architectural plans include symmetric multiprocessing (SMP) as well as multi-node, in-memory distributed processing, using a cluster of commodity servers.

CAES contains a cluster based on the Cloudera and Spark technology platform. It also contains a separate cluster based on the SAS-Viya distributed computing technology platform. CAES is directly connected to the data collection ecosystem through a variety of file and message-based interfaces that can bring inputs to and supply interventions from CAES in up to near real time.

CAES has been in operation since 2018 and has been used to execute several adaptive design models for research and in production. For the 2020 Decennial Census, CAES executed two critical models. First, the administrative record-based nonresponse follow-up (NRFU) model performed analyses on administrative data from various agencies such as the Internal Revenue Service, the Social Security Administration, and the Department of Housing and Urban Development (Morris, et al. 2015). This model identified cases in the NRFU workload that have existing administrative records and withheld such cases from field case workloads to avoid further contact attempts. Additionally, CAES supported the self-response quality assurance system (SRQA), an analytical model that performed analyses on administrative data received from various agencies to identify response data from the 2020 Census that may be fraudulent. CAES also hosts adaptive design models for various surveys such as the National Survey of College Graduates (NSCG; Coffey, et al. 2020), the Survey of Income and Program Participation (SIPP; Tolliver, et al. 2017), the National Teacher and Principal Survey (NTPS; Zotti 2019) and the Post-Enumeration Survey (PES). In the NSCG, for example, CAES has enabled the near-real time ingestion and processing of survey paradata and lightly edited response data to enable the

estimation of multiple Bayesian predictive models to inform optimization decision rules about how to allocate effort to cases (Coffey and Elliott 2022).

6.2.Survey Plus Environment

6.2.1. Systems for Web Scraping and Related Methods of Data Capture

In this new survey environment, an even broader array of data sources must be integrated. Automated methods like web scraping, web harvesting, or web data extraction are used to extract such data from websites. At BigSurv18 personnel from Statistics Netherlands presented a thorough look at the use of web-scraped data to accelerate and improve statistics (ten Bosch et al. 2018). Additionally, they highlighted various organizations and programs that were able to successfully explore the feasibility of integrating web-sourced data into their statistics. For example, Cavallo (2015) was able to compare measures of price-stickiness (i.e., the stability of the price of a product despite changes in cost, supply, or demand) from scraped data to measures from standard data sources including the Consumer Price Index (CPI) and scanner data. This measure required the collection of large volumes of micro-price data, which would have been impossible to scrape from the web without building a specialized program.

To enable similar research, in September 2020, the U.S. Census Bureau instituted a policy for the collection of web-sourced data. The policy was put in place to ensure that: (1) the data being collected are public information and our method of collection abides by any rules of access, terms of use, intellectual property rights of the data provider website; (2) the purpose of data collection is consistent with Census Bureau mission; and (3) the method of collection does not risk disclosure of confidential or legally protected data. To add the collection of such auxiliary

data to our ecosystem, Census established a multi-tenant webscraping platform, shown in Figure 1 (U.S. Census Bureau 2021a).

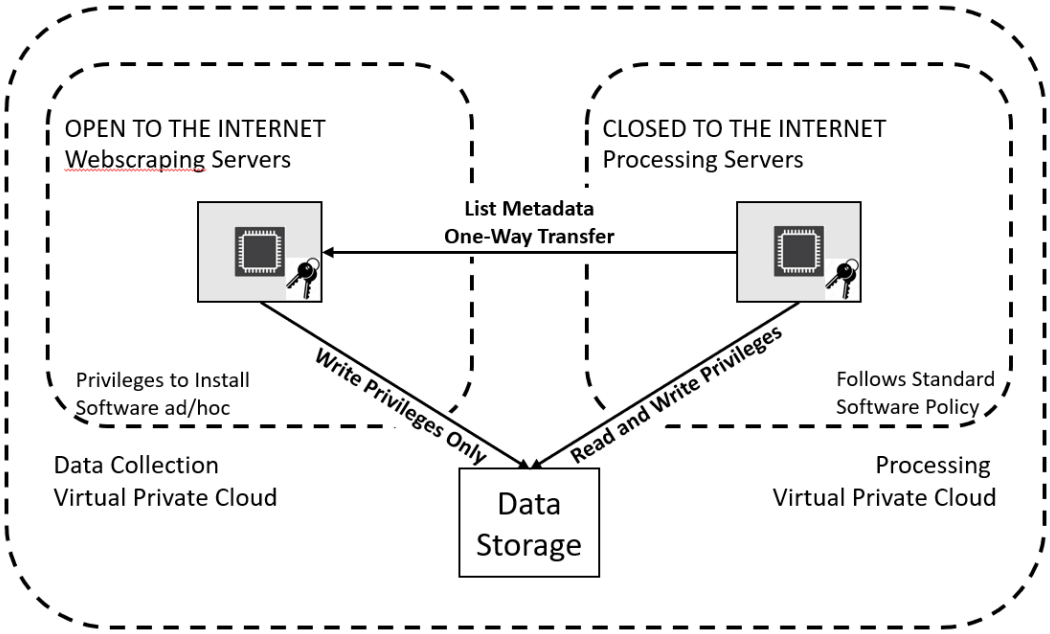

Figure 1. A Multi-Tenant Web-Scraping Platform

The webscraping platform has two distinct parts, namely: (i) servers open to the internet, and (ii) servers closed to the internet. The two parts of the platform exist isolated from each other except for a common data repository. Users must access the two parts of the platforms independently to accomplish distinct tasks.

The “open to the internet” servers, allow users complete access to the internet and users can therefore access any publicly available data on the internet. They also have access to, and therefore can install and use, the latest and emerging webscraping software tools and libraries available on the internet. Users can run tools on these servers to scrape and deposit data into the common data repository with “write-only” access.

The “closed to the internet” servers allow “read only” access to data that has been scraped and deposited into the common data repository. Here, users can only access Census Bureau standard modeling and analysis tools for “post scraping” processing of the data. Another one-way connection allows users to pass metadata such as a list of web sites to be scraped from the closed to the open part of the platform. Within this environment, the project team has access to the appropriate packages and can ensure compliance with bureau-wide disclosure guidelines. They also have access to a network of individuals building similar programs that encourages knowledge sharing.

This environment is still under development, but Census Bureau staff have already begun developing web scraping applications. For example, one ongoing project aims to scrape teacher names and subjects from school websites alongside information collected in the Teacher Follow-up Survey (TFS) of the National Teacher and Principal Survey (NTPS) (Mathur et al. 2021b). Successful work on this project could lead to imputation or other supplementation with information extracted from public school websites to portions of data collection operations.

6.2.2. General Systems for Data Capture and Integration

New platforms to acquire, process, integrate, and disseminate Census Bureau data assets will continue to be needed as new data sources and technologies come into use. This also affects how adaptive procedures will be carried out.

The need for the integration of acquisition, processing and dissemination ecosystems is one of the key drivers behind the development of an Enterprise Data Lake (EDL). As Figure 2 (Peters and Tracy 2020) shows, the EDL will allow for the integration of all acquired data including data collected from surveys, administrative and other external sources. The EDL will provide data to the adaptive design systems, as well as data for production of estimates for external publication.

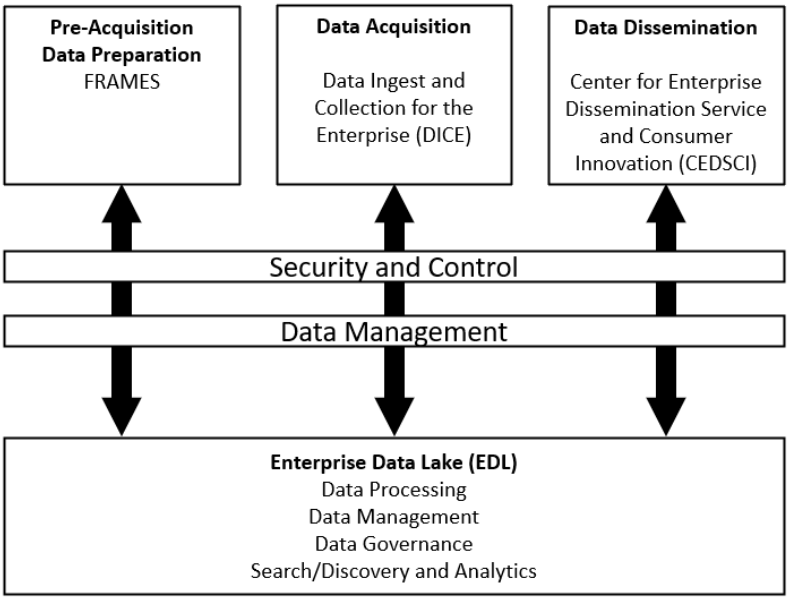

Figure 2. Schematic Design of the Enterprise Data Lake (EDL)

In the future, the EDL will not only be the reservoir of all acquired data, but it will also provide a cloud-based computation platform to process that data. This will allow CAES-like model execution capabilities to exist directly on the EDL.

Adaptive procedures that are informed by data and models on the EDL will be executed by the new Census OCS (nOCS) displayed in Figure 3 (U.S. Census Bureau 2021b; Mathur et al. 2021a; Mathur et al. 2021b), which will be the centerpiece of the Census data ecosystem.

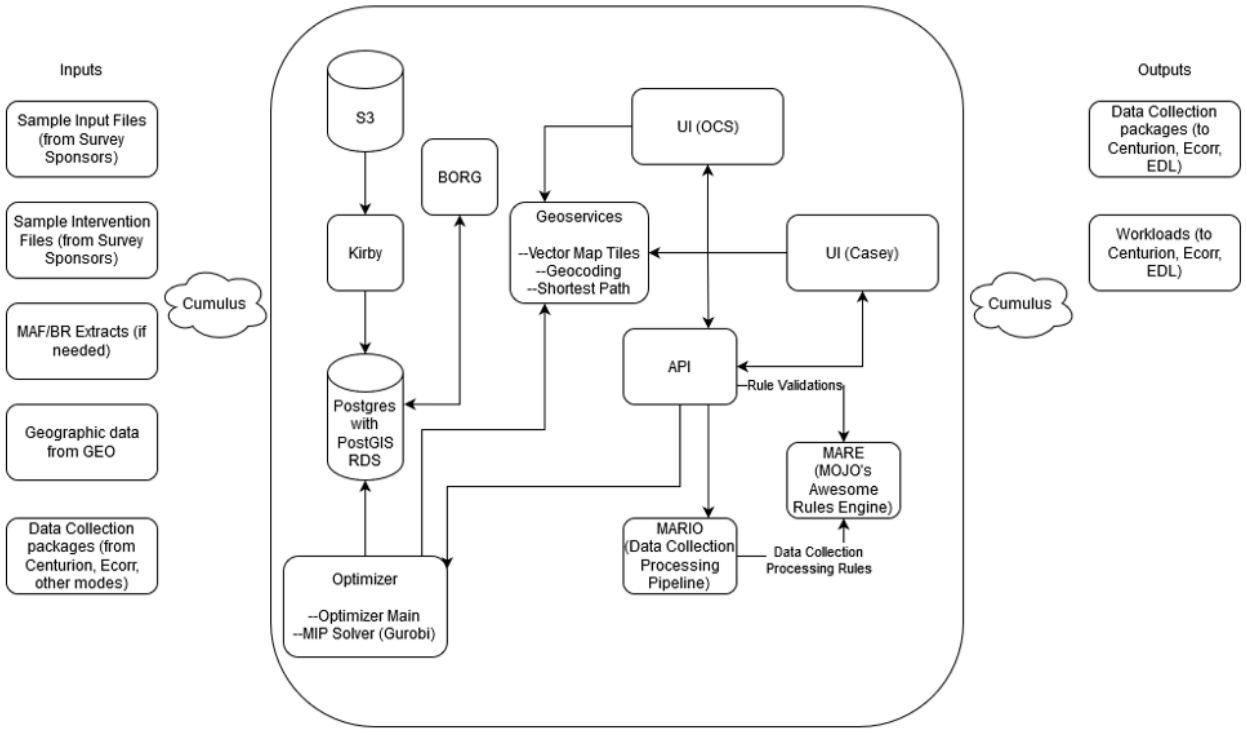

Figure 3. The New Census Operational Control System (nOCS)

There will be two ways for adaptive procedures and interventions to be executed in the nOCS. First, for complex models requiring CAES-like processing on the EDL, intervention messages will pass from the EDL to the nOCS through an event processing pipeline (called MARIO in Figure 3). Less complex adaptive interventions can be stored in the nOCS's internal rules engine (called MARE in Figure 3) that will allow Census survey sponsors to directly input specific intervention rules through a user interface. MARE will evaluate those business rules and process appropriate interventions within the nOCS.

As survey data collection and information production continues to expand to include new data sources, computational infrastructure needs will likely continue to evolve as well.

7. Adaptive Procedures: Quality, Cost, and Risk Profiles and Empirical Evaluation

As noted in previous sections, the overall goals of adaptive design involve improvement of the overall profiles of data quality, risk and cost, within the context defined by the overall societal and data-capture environment, and related operational constraints. This section provides a qualitative review of some of the environmental and operational factors that can be especially important for the "survey" and "survey plus" cases. In many applications, implementation of these ideas requires mathematical development and related empirical evaluations. In-depth exploration of these implementation issues will be beyond the scope of the current paper, but Appendix B provides a brief overview of some of underlying mathematical issues, with emphasis on distinctions between evaluation criteria that are conditional on, or averaged over, specific sources of random variability.

7.1. Current Environment

The evaluation of adaptive procedures is a critical component of adaptive design. Survey organizations have attempted to evaluate empirically the properties of their adaptive and responsive procedures, typically during survey data collection operations, but also through simulation. We summarize several recent empirical assessments here, but also see Tourangeau, et al. (2017) for a detailed review of select adaptive and responsive experiments prior to 2013. Measures of quality and cost vary across these studies, with quality measures including response rates, R-indicators, or retention rates (in longitudinal surveys). Similarly, costs are defined a variety of ways, including number of contact attempts, interviewer hours, or data collection costs.

Peytchev, et al. (2010) considered coefficients estimated from propensity models that relied upon historical survey data and used these historical estimates in conjunction with current covariates in a new survey sample in order to classify cases into propensity strata for incentivization.

However, the monetary incentives offered neither improved response rates nor reduced nonresponse bias for this particular study. For another study, Peytchev (2014) used paradata collected during a random digit dial (RDD) telephone survey to identify and stop work for cases falling below a pre-defined response propensity threshold. There was no statistically significant difference in the number of interviews obtained using this adaptive strategy versus the standard data collection protocol, though there was a statistically significant decrease in the mean number of call attempts needed to complete an interview, suggesting significant data collection cost savings.

Wagner, et al. (2012), described a complex case prioritization experiment carried out in the National Survey of Family Growth (NSFG), an in-person interviewer-administered survey. Cases could be prioritized based on variety of measures, including base weights, paradata from the early part of the NSFG, and characteristics collected in the screener portion of the survey. For most of the experimental prioritizations, measures of data quality (e.g., response rates, CVs of response rates, etc.) were improved when compared to the control group.

Tolliver, et al. (2019) discussed case prioritization in the Survey of Income and Program Participation (SIPP), a longitudinal, in-person, interviewer-administered survey. Cases could be prioritized by either a set of static business rules, or a dynamic prioritization that leveraged paradata accumulated during the survey period. The goal was to reduce attrition in the fourth wave of the survey, hopefully leading to increased representativeness of cases, measured by R-indicators. The authors found that the static business rules did not have a significant effect on R-indicators, but that the dynamic case prioritization led to statistically significantly higher R-indicators than in the control group.

Coffey et al. (2020) achieved statistically significant increases in representativeness, measured by R-indicators (Schouten, Shlomo and Skinner 2011), in the National Survey of College Graduates, a sequential multimode survey. This increase was achieved by increasing effort on some cases, while decreasing effort on others. Response rates and mean costs-per-case were consistent across the group managed using adaptive procedures and the control group. Recently, Coffey and Elliott (2022) demonstrated, through the use of an optimization rule, that data collection costs in the NSCG could be reduced by nearly 10% versus the control group, without significant decreases in unweighted response rates nor increases in root mean squared error (RMSE) of the mean of a single key survey estimate, self-reported salary.

In addition to empirical evaluations of some adaptive interventions, simulations have been used to demonstrate how more complex adaptive procedures might be implemented and what their expected effects might be. For example, Beaumont, Bocci and Haziza (2014) simulated case prioritization, based on the likelihood of a case to reduce the variance of survey estimates. Vandenplas, Loosveldt and Beullens (2017) also simulate adaptive interventions, here driven by fieldwork power, a productivity metric. The goal was to monitor interviewer productivity versus expected benchmarks to identify abnormalities in field data collection, allowing corrective interventions to be carried out in real time. Paiva and Reiter (2017) discussed a method for determining whether to continue data collection based on cost properties of the remaining non-respondent cases versus how their predicted responses could change estimates, imputing for non-respondents at a fixed point in time under a variety of different model assumptions. Lewis (2017) applied univariate tests to survey estimates after recruitment phases to identify whether to apply stopping rules to data collection operations.

Despite the promise of some adaptive procedures, it is difficult to generalize these findings across surveys due to wide variability of available interventions and the conditions under which interventions are carried out. The sheer variety of design features, including the length of data collection, modes available in a survey, and target population, combined with operational and environmental conditions such as those discussed in Section 3.1 make generalization difficult, continued experimentation critical, and widespread adoption slow.

In addition, some aspects of adaptive and responsive survey design have parallels in online and offline industrial quality control; and some related statistical issues arise in both settings. For example, in Evolutionary Operation (EVOP; Box and Draper 1969), manufacturing processes are improved in incremental ways over time, leading to improved product output. Some areas of EVOP use the working assumption that external environmental factors are largely static, and so improvements will have a persistent effect on production. This leads to process adaptations being integrated into the baseline production operations. This approach can require reconsideration for cases (including both industrial quality control and adaptive survey operations) for which important external factors are highly dynamic. For such applications (e.g., survey operations in which cases with response propensities under a set threshold receive a particular data collection feature), it can be problematic to produce a rigorous evaluation of the effectiveness of the intervention. Estimating the lasting effect of interventions based on historical experiments is a difficult problem, but an important one as data collection methodology continues to evolve.

7.2. Survey Plus Environment

In the new environment, it will continue to be important to evaluate the effect of interventions or changes to data collection procedures with respect to quality, cost and risk profiles. In addition,

however, it will also become important to evaluate empirically, on an ongoing basis, the quality, cost and risk that the use of alternative data sources bring to the information production process.

In particular, one broad source of risk with alternative data sources is that survey organizations do not "design" this data, nor have control over the data production process. This also means that we may not have control over, or even know, when processes for data production vary or when changes to those processes occur. For example, the National Ambulatory Medical Care Survey (NAMCS) attempted to mix information which was abstracted from medical records by interviewers with information directly extracted from electronic health records (EHR) to take advantage of the proliferation of EHRs for data collection. However, there were technological, analytic and disclosure challenges with EHR data (DeFrances and Lau 2018). Analytic issues arose for a variety of reasons around the assumptions that were made about the EHR production systems, such as assuming a record was static – this may be true for a paper record created by an interviewer abstracting data, but not for an electronic system that may overwrite or update information. Additionally, EHRs often had errors, or were missing fields required for the NAMCS reporting requirements.

Issues around the lack of control over the data production process may mean that more of the work associated with adaptive procedures will need to focus on the ongoing quality of alternative data streams, both for continuing periodic production (as occurs in surveys such as the American Community Survey (ACS) or the SIPP), and for the use of these data streams for one-off information production exercises. Thus, for some cases it will be important to have adaptive features in the procedures used to monitor the quality of the data provided by a specified alternative data source. Future evaluation methods may also require more focus on the production process for these alternative data sources, as well as additional psychological and

cognitive evaluations to better understand the conceptual overlap among alternative data sources, survey items, and the actual target estimand of interest. Given the fact that adaptive procedures in this new environment may not only affect data collection operations, but also sampling and estimation procedures, determining all aspects of alternative data that need to be evaluated is a critical need to utilizing these sources effectively. In some cases, the abovementioned evaluations may focus on general qualitative assessment. In other cases, however, it may be feasible to develop rigorous procedures to evaluate (under conditions) the conditional bias and conditional variance of some of the key estimates produced through the expanded suite of adaptive procedures. For example, one may develop estimators of conditional variances that account for the sources of variability associated with the imputation procedures described in Section 2.2; measurement error and other imperfections in specific non-survey data sources; and the effects of weighting used to combine information from multiple sources.

8. Discussion

In recent years, the production of high-quality statistical information has encountered extraordinary opportunities arising from the expanded availability of multiple data sources (e.g., administrative records and web-scraping, as well as customary sample surveys); and has also encountered important challenges arising from, e.g., declining survey response rates, as well as open questions about the quality, risk and cost profiles related to non-survey data sources. Consequently, it will be important to expand the concept of "design" to include the focused allocation of resources for production of high-quality statistical information on a sustainable and cost-effective basis; and to explore a wide range of ways in which one may adaptively improve the quality, risk and cost profiles of statistical information production procedures, based on paradata or other data that were not available at the start of the design process.

This paper provides an overview of dimensions for consideration, in order to extend customary "adaptive design" approaches to cases that use data from both sample surveys and non-survey data sources. Similar to typical adaptive survey designs, it is necessary to: (1) identify the goals of a particular information production process (survey or otherwise), (2) determine the design features available for adaptation, (3) identify available auxiliary data that could inform adaptation, (5) define decision rules to drive adaptation, (6) implement systems that enable adaptation, and finally, (7) develop measures by which to evaluate the impact of adaptation on the quality, cost and timeliness of the information production process.

While the steps to enable adaptation may be similar whether we consider a survey-focused or a survey-plus environment, focused research in a variety of areas is required to realize the benefits to information production of the integration of alternative data sources. Survey methodological research has a rich literature focused on topics such as the effectiveness of different frames for survey data quality; the impact on response from different modes of contact; the impact on measurement from different modes of data collection; and the impact on data quality from different imputation and weighting methods. In a survey-plus environment, it is necessary to extend those areas of research to evaluate the effectiveness of alternative data sources to meet estimation requirements related to coverage, accuracy, and timeliness; the impact on measurement from the linkage and use of alternative data sources; and the impact on data quality of different methods for integration, calibration, and use of alternative data sources. In particular, investigation of the quality and stability of alternative data sources will be critical, as will the further development of record linkage methodologies to enable the integration of multiple data sources for high quality information production.

In future work, there may be special interest in identification of particular application areas in which (i) paradata or other near-real-time process data may offer important insights into important dimensions of quality, risk or cost; (ii) those data point to practical adaptations (design modifications) that have the potential to produce substantial improvements in the overall profiles of quality, risk and cost, at least in some commonly encountered cases; and (iii) we have realistic methods to evaluate the resulting improvements, and to explain to key stakeholders the practical impact – and potential limitations – of the resulting adaptive procedures.

References:

Bates, N. 2009. "Cell Phone-Only Households: A Good Target for Internet Surveys?" *Survey Practice*, *2*(7), 2942.

Bates, N., Dahlhamer, J., Phipps, P., Safir, A., and L. Tan. 2010. "Assessing contact history paradata quality across several federal surveys." In *Proceedings of the Section on Survey Research Methods*, 91-105.

Beaumont, J.-F. 2020. "Are Probability Surveys Bound to Disappear for the Production of Official Statistics?" *Survey Methodology,* 46, 1-28.

Beaumont, J.-F., Bocci, C., and D. Haziza. 2014. "An adaptive data collection procedure for call prioritization." *Journal of Official Statistics*, 30, 607-621.

Bell, W., Basel, W., Cruse, C., Dalzell, L., Maples, J., O'Hara, B., and D. Powers. 2007. "Use of ACS Data to Produce SAIPE Mode-Based Estimates of Poverty for Counties." *Working Paper*. Center for Economic Studies, U.S. Census Bureau. (Available at: https://www.census.gov/library/working-papers/2007/demo/bell-01.html).

Benedetto, G., Motro, J., and M. Stinson. 2015. "Introducing Parametric Models and Administrative Records into 2014 SIPP Imputations." Federal Committee on Statistical Methodology (FCSM) Research Conference. Washington, DC. (Available at: https://nces.ed.gov/FCSM/pdf/D1_Benedetto_2015FCSM.pdf)

Biemer, P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly*, 74:5, 817-848, doi: 10.1093/poq/nfq058

Biemer, P., and A. Peytchev. 2012. “Census geocoding for nonresponse bias evaluation in telephone surveys: An assessment of the error properties.” *Public Opinion Quarterly*, 76(3). 432-452.

Biemer, P, deLeeuw, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L.E., Tucker, N.C., and B.T. West (Editors). 2017. *Total Survey Error in Practice,* New York: Wiley.

Blumberg, S.J. and J.V. Luke. 2021. “Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January-June 2021.” Technical report from the National Center for Health Statistics. (Available at: https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless202111.pdf).

Box, G.E. and N.R. Draper. 1969. *Evolutionary Operation: A Statistical Method for Process Improvement,* New York: Wiley.

Brackstone, G. 1999. “Managing Data Quality in a Statistical Agency.” *Survey Methodology,* 25, 139-149.

Bradley, V.C., Kuriwaki, S., and M. Isakov. 2021. “Unrepresentative big surveys significantly overestimated US vaccine uptake.” *Nature,* 600, 695–700. doi: 10.1038/s41586-021-04198-4

Bureau of Labor Statistics (BLS) 2021. “Effects of COVID-19 Pandemic and Response on the Consumer Expenditure Surveys.” Technical report from the Bureau of Labor Statistics. (Available at: https://www.bls.gov/covid19/effects-of-covid-19-pandemic-and-response-on-the-consumer-expenditure-surveys.htm).

Callegaro, M., Manfreda, K.L., and V. Vehovar. 2015. *Web Survey Methodology*, London: SAGE.

Cavallo, A. 2015. “Scraped Data and Sticky Prices.” Working paper from the *National Bureau of Economic Research (NBER).* (Available at: http://www.nber.org/papers/w21490).

Christen, P. 2019. Data Linkage: The Big Picture. *Harvard Data Science Review*, 1(2). https://doi.org/10.1162/99608f92.84deb5c4

Citro, C.F. 2014a. “From Multiple Modes for Surveys to Multiple Sources for Estimates.” *Survey Methodology*, 40, 137-161.

Citro, C.F. 2014b. “Principles and Practices for a Federal Statistical Agency: Why, What, and to What Effect.” *Statistics and Public Policy*, 1, 51-59, doi:10.1080/2330443X.2014.912953

Cochran, W.G. 1977. *Sampling Techniques, Third Edition*, New York: Wiley.

Coffey, S., Reist, B., and A. Zotti. 2015. “Static Adaptive Design in the NSCG: Results of Targeted Incentive Timing Study.” Presentation at the 2015 Joint Statistical Meetings, August 2015. (Available at: https://ww2.amstat.org/meetings/jsm/2015/onlineprogram/AbstractDetails.cfm?abstractid=317133).

Coffey, S., Reist, B., and P. Miller. 2020. “Interventions on Call: Dynamic Adaptive Design in the National Survey of College Graduates.” *Journal of Survey Statistics and Methodology*, 8:4, 726-747, doi:10.1093/jssam/smz026.

Coffey, S. and M.R. Elliott. 2022. “Optimizing Data Collection Interventions to Balance Cost and Quality in a Sequential Multimode Survey.” Manuscript in preparation.

Cornesse, C. 2020. “The utility of auxiliary data for survey response modeling: Evidence from the German Internet Panel.” *Survey Methods: Insights from the Field, Special Issue:*

'Fieldword Monitoring Strategies for Interviewer-Administered Surveys.'. (Available at: https://surveyinsights.org/?p=11849).

Couper, M.P. 2000. "Usability Evaluation of Computer-Assisted Survey Instruments." *Social Science Computer Review*, 18 (4), 384-396.

Couper, M.P. 2017. "Birth and Diffusion of the Concept of Paradata (in Japanese, translated by W. Matsumoto)." *Advances in Social Research*, 18, 14-26. (Available at: http://jasr.or.jp/english/JASR_Birth%20and%20Diffusion%20of%20the%20Concept%20of%20Paradata.pdf).

Dahlhamer, J. 2017. "Adaptive Design in the NHIS: Implementing Adaptive Design in the National Health Interview Survey: A Case Prioritization Experiment." Presentation at the 2017 AAPOR Conference. May 2017. New Orleans, LA.

DeFrances, C.J., and D.T. Lau. 2018. "Collecting Electronic Health Record Data for the National Ambulatory Medical Care Survey and the National Hospital Care Survey." *Proceedings from the 2018 Federal Committee on Statistical Methodology Conference*, Washington, DC, (Available at: https://nces.ed.gov/FCSM/pdf/J4_DeFrances_2018FCSM.pdf).

de Leeuw, E.D. 2005. "To mix or not to mix data collection modes in surveys." *Journal of Official Statistics*, 21(5). 233-255.

Dillman, D.A. 1978. *Mail and Telephone Surveys: The Total Design Method* (Vol. 19). New York: Wiley.

Elliott, M.R. and R. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science*, 32, 249-264.

Eltinge, J.L. 2013. “Integration of matrix sampling and multiple-frame methodology.” *Proceedings of the 59th World Statistical Congress*.)Available at: https://www.statistics.gov.hk/wsc/IPS033-P4-S.pdf).

Eltinge, J.L. 2018. Deming Award Lecture, Joint Statistical Meetings. Available through: JSM 2018 Plenary Session Webcasts (amstat.org)

FCSM. 2020. A Framework for Data Quality, *Working Paper FCSM-20-04*, Federal Committee on Statistical Methodology. (Available at: https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf).

Fuller, W.A. 1991. Regression Estimation in the Presence of Measurement Error. In *Measurement Errors in Surveys* (eds P.P. Biemer, R.M. Groves, L.E. Lyberg. N.A. Mathiowetz and S. Sudman), pp. 617-635. New York: Wiley.

Groves, R.M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

Giefer, K., Williams, A., Benedetto, G. and Joanna Moro. 2015. “Program Confusion in the 2014 SIPP: Using Administrative Records to Correct False Positive SSI Reports.” *Proceedings from the 2015 Federal Committee on Statistical Methodology Conference*, Washington, DC. (Available at: https://nces.ed.gov/fcsm/pdf/I1_Giefer_2015FCSM.pdf).

Groves, R.M. and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys.* New York: Wiley.

Groves, R. M. and S.G. Heeringa. 2006. “Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs.” *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

Groves, Robert M. and L.E. Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly*, 74:5, 849–879.

Hand, D.J. 2018. "Statistical Challenges of Administrative and Transaction Data." *Journal of the Royal Statistical Society, Series AI*, 181, 555-605.

Harter, R., Battaglia, M.P., Buskirk, T.D., Dillman, D.A., English, N, Fahimi, M. Frankel, M.R., Kennel, T., McMichael, J.P., McPhee, C.B., Montaquila, J. Yancey, T., and A.L. Zukerberg. 2016. *Report of the AAPOR Task Force on Address-Based Sampling.* (Available at: https://www.aapor.org/Education-Resources/Reports/Address-based-Sampling.aspx).

Jackson, M.T., McPhee, C. B., and P. Lavrakas. 2020. "Using Response Propensity Modeling to Allocate Noncontingent Incentives in an Address-Based Sample: Evidence from a National Experiment." *Journal of Survey Statistics and Methodology*, 8(2). 385–411, doi:10.1093/jssam/smz007

Larsen, L.J., Lineback, J.F., and B. Reist. 2020. "Continuing to Explore the Relation between Economic and Political Factors and Government Survey Refusal Rates: 1960–2015." *Journal of Official Statistics*, 36(3). 489-505, doi:10.2478/jos-2020-0026.

Lewis, T. 2017. "Univariate tests for phase capacity: tools for identifying when to modify a survey's data collection protocol", *Journal of Official Statistics*, 33(3). 601-624.

Lohr, S.L. 2021. "Multiple-Frame Surveys for a Multiple-Data-Source World." *Survey Methodology*, 47, 229-263. (Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-eng.pdf?st=AIzbgaSL).

Lohr, S.L., T.E. Raghunathan. 2017. "Combining Survey Data with Other Data Sources." *Statistical Science*, 32, 293-312.

Lohr, S., and J.N.K. Rao. 2006. "Estimation in Multiple-Frame Surveys." *Journal of the American Statistical Association*, *101*(475). 1019–1030. (Available at: http://www.jstor.org/stable/27590779).

Mathur, A., Castro, M., and S. Khaneja. 2021. "Automated Collection of Publicly Available Data from the Internet." Presentation at the *2021 Federal Committee on Statistical Methodology Conference*, Washington, DC. (Available at: https://copafs.org/wp-content/uploads/2021/11/C2Castro.pptx).

Mathur, A., Khaneja, S., and M. Minoo. 2021. "A New Tool to Supplement Survey Data Will Reduce Respondent Burden." Census Counterparts Dated 12/16/2021, Internal Unpublished Memo.

Meng, X. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox and the 2016 U.S. Presidential Election." *Annals of Applied Statistics*, 1-42. (Available at: https://statistics.fas.harvard.edu/files/statistics-2/files/statistical_paradises_and_paradoxes.pdf).

Morris, D.S., Keller, A., and B. Clark. 2015. "An Approach for Using Administrative Records to Reduce Contacts in the 2020 Census." *Working Paper DSSD-WP2015-06*. Decennial Statistical Studies Division, U.S. Census Bureau. (Available at: https://www.census.gov/library/working-papers/2015/dec/DSSD-WP2015-06.html).

Mule, T. 2021. "2020 Census: Administrative Record Usage." Presentation to the *National Academies of Science Panel to Evaluate the Quality of the 2020 Census*, Washington, DC. (Available at: https://www.nationalacademies.org/documents/embed/link/LF2255DA3DD1C41C0A42D3B

EF0989ACAECE3053A6A9B/file/DD09D62D4FCE1728687395A700936E6B3E50A9EC4D8B.)

National Academies of Sciences, Engineering, and Medicine (NASEM). 2017. Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. Washington, DC: The National Academies Press, doi: 10.17226/24893.

National Academies of Sciences, Engineering, and Medicine (NASEM). 2021. *Principles and Practices for a Federal Statistical Agency: Seventh Edition.* Washington, DC: The National Academies Press.

Neyman, J. 1938. "Contribution to the Theory of Sampling Human Populations." *Journal of the American Statistical Association*, 33:201, 101-116. doi: 10.1080/01621459.1938.10503378.

Olson, K., Wagner, J., and R. Anderson. 2021. "Survey Costs: Where are We and What is the Way Forward?" *Journal of Survey Statistics and Methodology*, 9(5). 921-942, doi: 10.1093/jssam/smaa014

Paiva, T., and J. Reiter. 2017. "Stop or continue data collection: a nonignorable missing data approach for continuous variables", *Journal of Official Statistics*, 33(3). 579-599.

Peters, D., and S. Tracy. 2020. "Census Enterprise Data Management (EDM) featuring the Enterprise Data Lake (EDL)." U.S. Census Bureau Unpublished Internal Technical Documentation.

Peytchev, A., Rosen, J., Riley, S. Murphy, J., and M. Lindblad. 2010. "Reduction of Nonresponse Bias through Case Prioritization." *Survey Research Methods*, 4, 21–29.

Peytchev, A. 2014. “Models and Interventions in Adaptive and Responsive Survey Designs.” DC-AAPOR Panel on Adaptive Survey Design. Washington, DC. (Available at: http://dc-aapor.org/ModelsInterventionsPeytchev.pdf).

Rao, J.N.K. 2021. “On making valid inferences by combining data from surveys and other sources.” *Sankhyā, Series B*, 83-B, 242–272.

Robinson, S. and K. Willyard. 2021. “Small Area Health Insurance Estimates: 2019.” *Working Paper P30-09*. Small Area Health Insurance Estimates Team, U.S. Census Bureau. (Available at: https://www.census.gov/content/dam/Census/library/publications/2021/demo/p30-09.pdf).

Rosenblum, M., Miller, P., Reist, B., Stuart, E., Thieme, M., and T. Louis. 2019. “Adaptive Design in Surveys and Clinical Trials: Similarities, Differences, and Opportunities for Cross-Fertilization.” *Journal of the Royal Statistical Society, Series A*, 182:3, 963-982.

Sarndal, C.-E., B. Swensson and J. Wretman. 1992. *Model-Assisted Survey Sampling*. New York: Springer.

Särndal, C.-E. and S. Lundström. 2008. “Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator.” *Journal of Official Statistics*, 24, 167-191.

Schouten, B., Cobben, F., and J. Bethlehem. 2009. “Indicators for the representativeness of survey response.” *Survey Methodology*, 35(1). 101-113.

Schouten, B., Shlomo, N., and C. Skinner. 2011. “Indicators for monitoring and improving representativeness of survey response.” *Journal of Official Statistics*, 27:2, 231-253.

Schouten, B., Peytchev, A., and J. Wagner. 2017. *Adaptive Survey Design*, Boca Raton, Florida: CRC Press.

Silvia, P.J., Kwapil, T.R., Eddington, K.M., and L.H. Brown. 2013. "Missed Beeps and Missing Data: Dispositional and Situational Predictors of Nonresponse in Experience Sampling Research." *Social Science Computer Review*, 31:4, 471-481, doi: 10.1177/0894439313479902

ten Bosch, O., Windmeijer, D., van Delden, A., and G. van den Heuvel. 2018. "Web scraping meets survey design: combining forces." Statistics Netherlands, The Hague, The Netherlands, October 25-27, 2018.

Thieme, M. and A. Mathur. 2014. "Designing and Architecting a Shared Platform for Adaptive Data Collection in Surveys and Censuses." *Proceedings of the 2014 Joint Statistical Meetings*, 3741-3751.

Thalji, L. Hill, C., Mitchell, S., Suresh, R., Speizer, H., and D. Pratt. 2013. "The General Survey System Initiative at RTI International: An Integrated System for the Collection and Management of Survey Data." *Journal of Official Statistics*, 29:1, 29-48.

Tolliver, K., Fields, J., Coffey, S., and B. Reist. 2017. "Prioritizing Cases Strategically for the Survey of Income and Program Participation (SIPP) Using R-indicator and other Business Rule Criteria." Presentation at 5th Workshop in Adaptive and Responsive Survey Design, August 2017.

Tolliver, K., Fields, J., Coffey, S., and A. Nagle. 2019. "Combatting Attrition Bias Using Case Prioritization in the Survey of Income and Program Participation." *Proceedings from the 2019 AAPOR Conference*, Toronto, Ontario. (Available at: www.asasrms.org/Proceedings/y2019/files/1199523.pdf).

Tourangeau, R.J., Brick, M., Lohr, S., Li, J. 2017. “Adaptive and Responsive Survey Designs: A Review and Assessment.” *Journal of the Royal Statistical Society, Series A*, 180, 203-223.

U.S. Census Bureau. 2021a. “Enterprise Webscraping Solution.” Unpublished Internal Technical Documentation.

U.S. Census Bureau. 2021b. “MOJO System Documentation.” Unpublished Internal Technical Documentation.

Valliant, R., Hubbard, F., Lee, S., and C. Chang. 2014. “Efficient use of commercial lists in US household sampling.” *Journal of Survey Statistics and Methodology*, 2(2). 182-209.

Vandenplas, C., Loosveldt, G., and K. Beullens. 2017. “Fieldwork monitoring for the European Social Survey: an Illustration with Belgium and the Czech Republic in Round 7.” *Journal of Official Statistics*, 33(3). 659-686.

van Berkel, K., van der Doef, S., and B. Schouten. 2020. “Implementing Adaptive Survey Design with an Application to the Dutch Health Survey.” *Journal of Official Statistics* (JOS). 36(3), 609-629.

Wagner, J., B. West, N. Kirgis, J. Lepkowski, W. Axinn, and S. Ndiaye. 2012. “Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection.” *Journal of Official Statistics*, 28(4). 477–499.

Wagner, James. 2019. “Estimation of Survey Cost Parameters Using Paradata.” *Survey Practice* 12 (1). doi:10.29115/SP-2018-0036.

Wagner, J., West, B.T., Coffey, S.M., and M.R. Elliott. 2020. “Comparing the Ability of Regression Modeling and Bayesian Additive Regression Trees to Predict Costs in a

Responsive Survey Design Context." *Journal of Official Statistics*, 36(4). 907-931, doi: 10.2478/jos-2020-0043.

Wagner, J., X. Zhang, M.R. Elliott, B.T. West, and S. Coffey. Under review. "An Experimental Evaluation of a Stopping Rule Aimed at Maximizing Cost-Quality Tradeoffs." *Submitted to the Journal of the Royal Statistical Society Series-A, December 2021*.

Walejko, G. and J. Wagner. 2018. "A Study of Interviewer Compliance in 2013 and 2014 Census Test Adaptive Designs." *Journal of Official Statistics*, 34:3, 649-670. doi:10.2478/jos-2018-0031

West, B. T., and F. Kreuter. 2013. "Factors affecting the accuracy of interviewer observations: Evidence from the National Survey of Family Growth." *Public Opinion Quarterly*, 77(2). 522-548.

West, B.T., Wagner, J., Coffey, S., and M.R. Elliott. 2021. "Deriving Priors for Bayesian Prediction of Daily Response Propensity in Responsive Survey Design: Historical Data Analysis vs. Literature Review." *Journal of Survey Statistics and Methodology*. In press. doi: https://doi.org/101093/jssam/smab036

Zotti, A. 2019. "Using Predictive Models to Assign Treatment Groups for NTPS 2017-18 Teacher Incentives Experiment." Presentation at 2019 Federal Computer Assisted Survey Information Collection Workshops, Washington, DC, Available at: https://www.census.gov/fedcasic/fc2019/ppt/4AZotti.pdf.

Appendix A: Illustrative Examples of Levels of Design Specification

Section 3 outlined the principal design features of interest for this paper, and noted that for both survey and non-survey data sources, decisions often will involve multiple levels of design specifications. Each level can potentially involve options that are adaptive in the broad sense that they may use data that were not available at the start of design work. Some of these cases have been established in survey practice for decades, while others have developed more recently. Four examples are as follows. Similar ideas also may apply to other cases that will not be explored in detail here, e.g., address-based sampling (e.g., Harter et al., 2016, and references cited therein) and extensions of multiple-frame, multiple-mode survey methods (e.g., Lohr and Rao, 2006; Eltinge, 2013; Citro, 2014a; Lohr, 2021; and references cited therein) to the integration of survey and non-survey data sources.

Example A.1: Customary Stratified Multistage Sample Surveys

Level 1: The (sub)population(s) and estimands of principal interest; the performance criteria (quality, risk and cost) important for production of the resulting estimates; operating constraints determined by the legal, regulatory and managerial environment; the stakeholder groups (along with applicable utility functions) with principal interest in the specified estimands and the resulting published estimates; determination of the overall availability of resources for the full trajectory of research, development and operations; and allocation of those resources to specific activities. The latter items include all applicable categories of resources, e.g., funding; specific data sources; technological and methodological capabilities; scarce skill sets; and related institutional capabilities.

Adaptive options for Level 1:

Appendix A: Illustrative Examples of Levels of Design Specification

For longstanding data production programs, the following may be viewed as adaptive, in the sense that changes in these areas were not necessarily anticipated when certain fundamental design features were established. One notable illustrative example involves model-based small domain estimation supplements, based on a "core" collection program (e.g., the U.S. Current Population Survey or the U.S. American Community Survey) that remains largely unchanged (Bell, et al. 2007; Robinson and Willyard 2021). For such cases, we need to consider both adjustments in the estimands of principal interest, based on changing needs of core stakeholder groups as well as adjustments in the set of stakeholders (and thus their utility functions) that receive principal attention in decisions on trade-offs among competing dimensions of quality, risk and cost.

Level 1 adaptations also include adjustments in the abovementioned operating constraints; in the overall amounts of resources available for the proposed work; and in allocation of those resources to specific tasks. In some cases, these Level 1 adaptations may involve allocation of substantial resources to evaluation of methodological properties of the proposed procedures within a specified operating environment. See, e.g., Fuller (1991) for discussion of resource allocation for assessment of measurement error magnitudes; and Groves and Couper (1998) for discussion of "designing for nonresponse."

Level 2: The conceptual strata, primary/secondary/ultimate sample units; frame(s) for those units; and selection probabilities at each stage.

Adaptive options for Level 2: Traditional multi-phase sampling options (e.g., Sarndal, Swensson and Wretman, 1992), as well as reductions in sample sizes for an ongoing survey (e.g., resulting

from budgetary constraints not known at the time of the initial multi-year sample selection) could be considered.

Level 3: Instrument(s) and fieldwork procedures.

Adaptive options for Level 3:

Mode assignment, as well as nonresponse follow-up strategies such as number of callbacks, special persuasion efforts, and basic question procedures could all be considered.

Level 4: Microdata edit, imputation and weighting procedures

Adaptive options for Level 4:

Coarsening of weighting or imputation cells, as well as data-driven decisions on edit and outlier detection/mitigation procedures could be considered.

Level 5: Variance estimation and inference

Adaptive options for Level 5:

Data-driven stratum collapse procedures could be considered.

Example A.2: Appending Administrative Record Data to Sample Units (cf. the "Surveys First" option at Statistics Canada)

Adaptive modification of Levels 3 (fieldwork) and 4 (imputation) from Example A.1: Use administrative record data to produce (a) improved response propensity models; and (b) imputation of key response items. Then focus nonresponse follow-up efforts on the sample units

for which the models from (a) indicate that the unit has a relatively high probability of response based on a specific follow-up procedure; and the models from (b) indicate relatively weak imputation performance for the unit, based, e.g., lack-of-fit diagnostics from the imputation model, or a large residual imputation variance, conditional on the predictor variables available for that nonresponding unit.

Example A.3: Using Sample Surveys to "Bridge the Gaps" in Available Administrative Records (cf. the "Administrative Records First" option at Statistics Canada)

Level 1: Still ultimately focused on the same estimands as in Example A.1. However, as an intermediate step, the sample design as such will focus on estimation of:

(a) the means of subpopulations that are not covered by the administrative records;

(b) the means of subpopulations formed by the intersection of two or more administrative record sources, in an extension of standard multiple-frame methodology; and

(c) estimation of regression coefficients needed to calibrate the relationship between the concepts aligned with the idealized estimands, and the measures available from a specific administrative record source

Levels 2-5: Modification of designs to improve the performance for production focused on the estimands specified in Level 1 could be considered.

Example A.4: Capture and Use of Web-Scraped Data

Appendix A: Illustrative Examples of Levels of Design Specification

Level 1: Legal, regulatory and contractual structures that may limit the form and extent of web-scaping that a statistical organization may perform within specified classes of public information

Adaptive features: Adjustments based on empirical results indicating that to do certain web-scraping tasks, we are encountering legal, regulatory or contractual issues not anticipated originally. One illustrative example arises from the perception that extensive web-scraping constitutes a denial-of-service attack and thus leads to countermeasures by the web host, which in turn leads to negotiations between the data collector and the web host.

Level 2: Consistency and reliability of the web sources

Adaptive features: Similar to the abovementioned approaches to the use of administrative records, we consider modification of designs to focus on the performance in production for estimands specified in Level 1. When using web sourced data, we also must consider the consistency and reliability (ten Bosch et al. 2018) of the data. Unlike many more traditioir data. That could mean that a data source found on the web one year might not be there to use again next year, making it difficult to rely on web sourced data to supplement data collection. Survey sponsors may need to consider adaptive procedures that allow for fluid changes in data sources in the event that web sources become unavailable or unreliable.

Level 3: Quality and format of the web sources

Adaptive Features: In addition to the availability of web sourced data, in-depth evaluations of the quality and usability of the web sourced data (Mathur, Khaneja, Minoo 2020) often are required before one makes a decision about whether, and how, to use those data.

Level 4: Coverage of the web sources

Appendix A: Illustrative Examples of Levels of Design Specification

Adaptive Features: When sourcing data across a large number of websites, it might become clear that quality data are available only for select subsets of the population of interest. This will require either an alternative form of data collection, or update weighting and imputation methods to make up for the gaps in the web sourced data.

Appendix B: Prospective Impact of Adaptive Procedures on Conditional and Unconditional Performance Profiles

The main sections of the paper discussed adaptive design, and related extensions to the integration of multiple data sources, in relatively qualitative and descriptive forms. Further development of these ideas may benefit from the use of some structured mathematical material, especially for evaluation of the conditional and unconditional properties of adaptive procedures.

B.1. Notation: Environmental, Substantive, Process and Design Variables

In an extension of ideas developed in Eltinge (2018), the following notation provides some mathematical structure for the discussion of properties of broad classes of adaptive procedures in survey-plus contexts.

First, we define $Z =$ Environmental variables (observed, uncontrolled)

Second, we consider a vector of substantive variables $Y = (Y_1, Y_2, Y_3)$ collected at the unit level, where

Y_1 = Outcome variables of principal interest

Y_2 = Predictor variables considered potentially important in modeling of substantive or process-related phenomena

Y_3 = Paradata considered potentially important for modeling of the data-capture or production process, e.g., predictors for use in models of propensity to respond to a survey request; of

propensity to consent to having records linked to survey responses; or microdata quality for either survey or non-survey data sources

Third, we have a design vector that describes the full set of resource allocation decisions $X = (X_1, X_2)$, including the subvectors

X_1 = Design features that are fully prespecified

$X_2 = X_2(Y_3; \gamma)$ = Design features that depend on preliminary paradata Y_3 initially collected from the sample units; and on idealized knowledge of a vector γ of process-model parameters. This may include dependencies at the individual unit level, or at coarser levels (e.g., neighborhoods or business firms)

$\hat{X}_2 = X_2(Y_3; \hat{\gamma})$ = Approximation to the idealized X_2 based on use of paradata Y_3 and the vector $\hat{\gamma}$ of estimated process-model parameters

In addition, each of X_1 and X_2 have sub-vectors describing resource-allocation decisions related to data sources, methodology, technology systems and administrative activities, respectively. Formally, for $j = 1, 2$: $X_j = (X_{j,Source}, X_{j,Method}, X_{j,System}, X_{j,Admin})$

Appendix B: Prospective Impact of Adaptive Procedures on Conditional and Unconditional Performance Profiles

B.2. Schematic Models for Performance Profiles: Quality, Risk and Cost

Design work (including adaptive and responsive design) is generally intended to provide a good balance of multiple dimensions of quality, risk and cost within the context defined by numerous environmental factors, operating constraints and limits on available information. Section 2 summarized some of the applicable dimensions of quality, risk and cost that have received principal attention in the literature to date.

A schematic model for the resulting "performance profile" vector is:

$$P = (Quality,\ Risk,\ Cost) = g_\theta(X, Z;\ \gamma) + e$$

where e is a vector of residual effects (uncontrolled and unobserved, with mean equal to zero); and γ is a vector of parameters for the performance profile and related dispersion effects arising from e.

Note that in the general case, the performance profile P is a random vector, with sources of random variability including the general residual term e; the observed but uncontrolled environmental vector Z; and the effects of paradata usage in the data-driven adaptations reflected in $\hat{X}_2 = X_2(Y_3; \hat{\gamma})$. Consequently, design decisions may benefit from consideration of several complementary approaches to evaluation of the properties of P, including:

(a) If we want to understand P in the context of our current environmental conditions Z, but averaging over the unobservable residuals e and treating the full design vector X as fixed:

$$E_e(P \mid Z) = g_\theta(X, Z;\ \gamma)$$

Note that the conditional expectation is itself a random vector, with its variability arising from the effects of the environmental factors Z.

(b) If we want to understand P , averaging over the distributions of both the environmental conditions Z and the residuals e, and treating the full design vector X as fixed:

$$E_Z\{E_e(P \mid Z)\} = E_Z\{g_\theta(X, Z; \gamma)\}$$

(c) If we want to understand P in the context of our current environmental conditions Z, but averaging over the unobservable residuals e and accounting for random variability arising from data-driven adaptation of some design features through use of $\hat{\gamma}$ and $\hat{X}_2$:

$$E_{Y_3}\{E_e(P \mid Z)\} = E_{Y_3}\left\{g_\theta\left((X_1, \hat{X}_2), Z; \hat{\gamma}\right)\right\}$$

(d) If we want to understand P , averaging over the distributions of the environmental conditions Z and residuals e, and also accounting for random variability arising from data-driven adaptation of some design features through use of $\hat{\gamma}$ and $\hat{X}_2$:

$$E_Z\left[E_{Y_3}\{E_e(P \mid Z)\}\right] = E_Z\left[E_{Y_3}\left\{g_\theta\left((X_1, \hat{X}_2), Z; \hat{\gamma}\right)\right\}\right]$$